"Learning Bayesian Networks."
Richard E. Neapolitan. Prentice Hall Series in AI, 2004.

how this means we often need to ascertain far fewer values than if we had to determine all values in the joint distribution directly. Before proving it, we illustrate what it means for a joint distribution to equal the product of its conditional distributions of all nodes given values of their parents in a DAG $\mathbb{G}$. This would be the case for a joint probability distribution P of the variables in the DAG in Figure 1.4 if, for all values of f, c, b, l, and h,

$$P(f,c,b,l,h) = P(f|b,l)P(c|l)P(b|h)P(l|h)P(h), \tag{1.7}$$

whenever the conditional probabilities on the right exist. Notice that if one of them does not exist for some combination of the values of the variables, then $P(b,l) = 0$ or $P(l) = 0$ or $P(h) = 0$, which implies $P(f,c,b,l,h) = 0$ for that combination of values. However, there are cases in which $P(f,c,b,l,h) = 0$ and the conditional probabilities still exist. For example, this would be the case if all the conditional probabilities on the right existed and $P(f|b,l) = 0$ for some combination of values of f, b, and l. So Equality 1.7 must hold for all nonzero values of the joint probability distribution plus some zero values.

We now state the theorem:

Theorem 1.4 *If $(\mathbb{G}, P)$ satisfies the Markov condition, then P is equal to the product of its conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist.*

Proof. *We prove the case where P is discrete. Order the nodes so that if Y is a descendent of Z, then Y follows Z in the ordering. Such an ordering is called an* ***ancestral ordering****. Examples of such an ordering for the DAG in Figure 1.4 are $[H, L, B, C, F]$ and $[H, B, L, F, C]$. Let $X_1, X_2, \ldots, X_n$ be the resultant ordering. For a given set of values of $x_1, x_2, \ldots, x_n$, let pa_i be the subset of these values containing the values of X_i's parents. We need to show that whenever $P(\mathsf{pa}_i) \neq 0$ for $1 \leq i \leq n$,*

$$P(x_n, x_{n-1}, \ldots, x_1) = P(x_n|\mathsf{pa}_n)P(x_{n-1}|\mathsf{pa}_{n-1}) \cdots P(x_1|\mathsf{pa}_1).$$

We show this by using induction on the number of variables in the network. Assume, for some combination of values of the x_i's, that $P(\mathsf{pa}_i) \neq 0$ for $1 \leq i \leq n$.

INDUCTION BASE: *Since PA_1 is empty,*

$$P(x_1) = P(x_1|\mathsf{pa}_1).$$

INDUCTION HYPOTHESIS: *Suppose for this combination of values of the x_i's that*

$$P(x_i, x_{i-1}, \ldots, x_1) = P(x_i|\mathsf{pa}_i)P(x_{i-1}|\mathsf{pa}_{i-1}) \cdots P(x_1|\mathsf{pa}_1).$$

INDUCTION STEP: *We need to show for this combination of values of the x_i's that*

$$P(x_{i+1}, x_i, \ldots, x_1) = P(x_{i+1}|\mathsf{pa}_{i+1})P(x_i|\mathsf{pa}_i) \cdots P(x_1|\mathsf{pa}_1). \tag{1.8}$$

There are two cases:

CASE 1: *For this combination of values*

$$P(x_i, x_{i-1}, \ldots, x_1) = 0. \tag{1.9}$$

Clearly, Equality 1.9 implies

$$P(x_{i+1}, x_i, \ldots, x_1) = 0.$$

Furthermore, due to Equality 1.9 and the induction hypothesis, there is some k, where $1 \leq k \leq i$, such that $P(x_k|\mathsf{pa}_k) = 0$. So Equality 1.8 holds.

CASE 2: *For this combination of values*

$$P(x_i, x_{i-1}, \ldots, x_1) \neq 0.$$

In this case,

$$\begin{aligned} P(x_{i+1}, x_i, \ldots, x_1) &= P(x_{i+1}|x_i, \ldots, x_1)P(x_i, \ldots, x_1) \\ &= P(x_{i+1}|\mathsf{pa}_{i+1})P(x_i, \ldots, x_1) \\ &= P(x_{i+1}|\mathsf{pa}_{i+1})P(x_i|\mathsf{pa}_i)\cdots P(x_1|\mathsf{pa}_1). \end{aligned}$$

The first equality is due to the rule for conditional probability, the second is due to the Markov condition and the fact that $X_1, \ldots X_i$ are all nondescendents of X_{i+1}, and the last is due to the induction hypothesis. ■

Example 1.31 *Recall that the joint probability distribution in Example 1.29 satisfies the Markov condition with the DAG in Figure 1.3 (a). Therefore, owing to Theorem 1.4,*

$$P(v, s, c) = P(v|c)P(s|c)p(c), \tag{1.10}$$

and we need only determine the conditional distributions on the right in Equality 1.10 to uniquely determine the values in the joint distribution. We illustrate that this is the case for $v1$, $s1$, and $c1$:

$$P(v1, s1, c1) = P(\mathsf{One} \cap \mathsf{Square} \cap \mathsf{Black}) = \frac{2}{13}$$

$$\begin{aligned} P(v1|c1)P(s1|c1)P(c1) &= P(\mathsf{One}|\mathsf{Black}) \times P(\mathsf{Square}|\mathsf{Black}) \times P(\mathsf{Black}) \\ &= \frac{1}{3} \times \frac{2}{3} \times \frac{9}{13} = \frac{2}{13}. \end{aligned}$$

Figure 1.5 shows the DAG along with the conditional distributions.

The joint probability distribution in Example 1.29 also satisfies the Markov condition with the DAGs in Figures 1.3 (b) and (c). Therefore, the probability

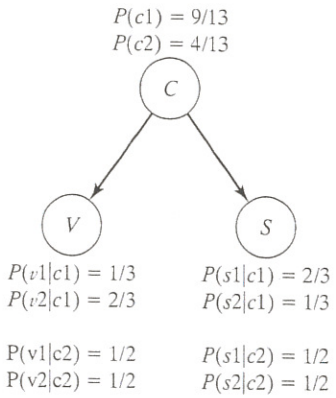

Figure 1.5: The probability distribution discussed in Example 1.31 is equal to the product of these conditional distributions.

distribution in that example equals the product of the conditional distributions for each of them. You should verify this directly.

Theorem 1.4 often enables us to reduce the problem of determining a huge number of probability values to that of determining relatively few. The number of values in the joint distribution is exponential in terms of the number of variables. However, each of these values is uniquely determined by the conditional distributions (due to the theorem), and if each node in the DAG does not have too many children, there are not many values in these distributions. For example, if each variable has two possible values and each node has at most one parent, we would need to ascertain less than $2n$ probability values to determine the conditional distributions when the DAG contains n nodes. On the other hand, we would need to ascertain $2^n - 1$ values to determine the joint probability distribution directly. In general, if each variable has two possible values and each node has at most k parents, we need to ascertain less than $2^k n$ values to determine the conditional distributions. So if k is not large, we have a manageable number of values.

Something may seem amiss to you. In Example 1.29, we started with an underlying sample space and probability function, specified some random variables, and showed that if P is the probability distribution of these variables and $\mathbb{G}$ is the DAG in Figure 1.3 (a), then $(P, \mathbb{G})$ satisfies the Markov condition. We can therefore apply Theorem 1.4 to conclude we need only determine the conditional distributions of the variables for that DAG to find any value in the joint distribution. We illustrated this in Example 1.31. However, as discussed in Section 1.3, in application we do not ordinarily specify an underlying sample space and probability function from which we can compute conditional distributions. Rather, we identify random variables and values in conditional distributions directly. For example, in an application involving the diagnosis of lung cancer, we identify variables like $SmokingHistory$, $LungCancer$, and $ChestXray$, and probabilities such as $P(SmokingHistory =$

yes), $P(LungCancer = present|SmokingHistory = yes)$, and $P(ChestXray = positive|\ LungCancer = present)$. How do we know the product of these conditional distributions is a joint distribution at all, much less one satisfying the Markov condition with some DAG? Theorem 1.4 tells us only that if we start with a joint distribution satisfying the Markov condition with some DAG, the values in that joint distribution will be given by the product of the conditional distributions. However, we must work in reverse. We must start with the conditional distributions and then be able to conclude the product of these distributions is a joint distribution satisfying the Markov condition with some DAG. The theorem that follows enables us to do just that.

Theorem 1.5 *Let a DAG $\mathbb{G}$ be given in which each node is a random variable, and let a discrete conditional probability distribution of each node given values of its parents in $\mathbb{G}$ be specified. Then the product of these conditional distributions yields a joint probability distribution P of the variables, and $(\mathbb{G}, P)$ satisfies the Markov condition.*

Proof. *Order the nodes according to an ancestral ordering. Let $X_1, X_2, \ldots, X_n$ be the resultant ordering. Next define*

$$P(x_1, x_2, \ldots, x_n) = P(x_n|\mathsf{pa}_n)P(x_{n-1}|\mathsf{pa}_{n-1}) \cdots P(x_2|\mathsf{pa}_2)P(x_1|\mathsf{pa}_1),$$

where PA_i is the set of parents of X_i in $\mathbb{G}$ and $P(x_i|\mathsf{pa}_i)$ is the specified conditional probability distribution. First we show this does indeed yield a joint probability distribution. Clearly, $0 \leq P(x_1, x_2, \ldots x_n) \leq 1$ for all values of the variables. Therefore, to show we have a joint distribution, Definition 1.8 and Theorem 1.3 imply we need only show that the sum of $P(x_1, x_2, \ldots x_n)$, as the variables range through all their possible values, is equal to one. To that end,

$$\begin{aligned}
&\sum_{x_1}\sum_{x_2}\cdots\sum_{x_{n-1}}\sum_{x_n} P(x_1, x_2, \ldots, x_n) \\
&= \sum_{x_1}\sum_{x_2}\cdots\sum_{x_{n-1}}\sum_{x_n} P(x_n|\mathsf{pa}_n)P(x_{n-1}|\mathsf{pa}_{n-1}) \cdots P(x_2|\mathsf{pa}_2)P(x_1|\mathsf{pa}_1) \\
&= \sum_{x_1}\left[\sum_{x_2}\left[\cdots\sum_{x_{n-1}}\left[\sum_{x_n} P(x_n|\mathsf{pa}_n)\right]P(x_{n-1}|\mathsf{pa}_{n-1})\cdots\right]P(x_2|\mathsf{pa}_2)\right]P(x_1|\mathsf{pa}_1) \\
&= \sum_{x_1}\left[\sum_{x_2}\left[\cdots\sum_{x_{n-1}}[1]P(x_{n-1}|\mathsf{pa}_{n-1})\cdots\right]P(x_2|\mathsf{pa}_2)\right]P(x_1|\mathsf{pa}_1) \\
&= \sum_{x_1}\left[\sum_{x_2}[\cdots 1\cdots]P(x_2|\mathsf{pa}_2)\right]P(x_1|\mathsf{pa}_1) \\
&= \sum_{x_1}[1]\,P(x_1|\mathsf{pa}_1) = 1.
\end{aligned}$$

It is left as an exercise to show that the specified conditional distributions are the conditional distributions they notationally represent in the joint distribution.

Finally, we show the Markov condition is satisfied. To do this, we need to show, for $1 \leq k \leq n$, *that whenever* $P(\mathsf{pa}_k) \neq 0$, *if* $P(\mathsf{nd}_k|\mathsf{pa}_k) \neq 0$ *and* $P(x_k|\mathsf{pa}_k) \neq 0$ *then* $P(x_k|\mathsf{nd}_k, \mathsf{pa}_k) = P(x_k|\mathsf{pa}_k)$, *where* ND_k *is the set of nondescendents of* X_k *in* $\mathbb{G}$. *Since* $\mathsf{PA}_k \subseteq \mathsf{ND}_k$, *we need only show* $P(x_k|\mathsf{nd}_k) = P(x_k|\mathsf{pa}_k)$.

First, for a given k, *order the nodes so that all and only nondescendents of* X_k *precede* X_k *in the ordering. Note that this ordering depends on* k, *whereas the ordering in the first part of the proof does not. Clearly, then,*

$$\mathsf{ND}_k = \{X_1, X_2, \ldots, X_{k-1}\}.$$

Let

$$\mathsf{D}_k = \{X_{k+1}, X_{k+2}, \ldots, X_n\}.$$

In what follows, $\sum_{\mathsf{d}_k}$ *means the sum as the variables in* d_k *go through all their possible values. Furthermore, notation such as* $\hat{x}_k$ *means the variable has a particular value; notation such as* $\hat{\mathsf{nd}}_k$ *means all variables in the set have particular values; and notation such as* pa_n *means some variables in the set may not have particular values. We have*

$$\begin{aligned}
P(\hat{x}_k|\hat{\mathsf{nd}}_k) &= \frac{P(\hat{x}_k, \hat{\mathsf{nd}}_k)}{P(\hat{\mathsf{nd}}_k)} \\
&= \frac{\sum_{\mathsf{d}_k} P(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k, x_{k+1}, \ldots, x_n)}{\sum_{\mathsf{d}_k \cup \{x_k\}} P(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{k-1}, x_k, \ldots, x_n)} \\
&= \frac{\sum_{\mathsf{d}_k} P(x_n|\mathsf{pa}_n) \cdots P(x_{k+1}|\mathsf{pa}_{k+1}) P(\hat{x}_k|\hat{\mathsf{pa}}_k) \cdots P(\hat{x}_1|\hat{\mathsf{pa}}_1)}{\sum_{\mathsf{d}_k \cup \{x_k\}} P(x_n|\mathsf{pa}_n) \cdots P(x_k|\mathsf{pa}_k) P(\hat{x}_{k-1}|\hat{\mathsf{pa}}_{k-1}) \cdots P(\hat{x}_1|\hat{\mathsf{pa}}_1)} \\
&= \frac{P(\hat{x}_k|\hat{\mathsf{pa}}_k) \cdots P(\hat{x}_1|\hat{\mathsf{pa}}_1) \sum_{\mathsf{d}_k} P(x_n|\mathsf{pa}_n) \cdots P(x_{k+1}|\mathsf{pa}_{k+1})}{P(\hat{x}_{k-1}|\hat{\mathsf{pa}}_{k-1}) \cdots P(\hat{x}_1|\hat{\mathsf{pa}}_1) \sum_{\mathsf{d}_k \cup \{x_k\}} P(x_n|\mathsf{pa}_n) \cdots P(x_k|\mathsf{pa}_k)} \\
&= \frac{P(\hat{x}_k|\hat{\mathsf{pa}}_k)\,[1]}{[1]} = P(\hat{x}_k|\hat{\mathsf{pa}}_k).
\end{aligned}$$

In the second-to-last step, the sums are each equal to unity for the following reason. Each is a sum of a product of conditional probability distributions specified for a DAG. In the case of the numerator, that DAG is the subgraph of our original DAG $\mathbb{G}$, *consisting of the variables in* D_k, *and in the case of the denominator, it is the subgraph consisting of the variables in* $\mathsf{D}_k \cup \{X_k\}$. *Therefore, the fact that each sum equals unity follows from the first part of this proof.* ■

Notice that the theorem requires that specified conditional distributions be discrete. Often in the case of continuous distributions it still holds. For example, it holds for the Gaussian distributions introduced in Section 4.1.3. However, in